Regression_Basics

Bird

1/27/2022

```
library(ggplot2)
```

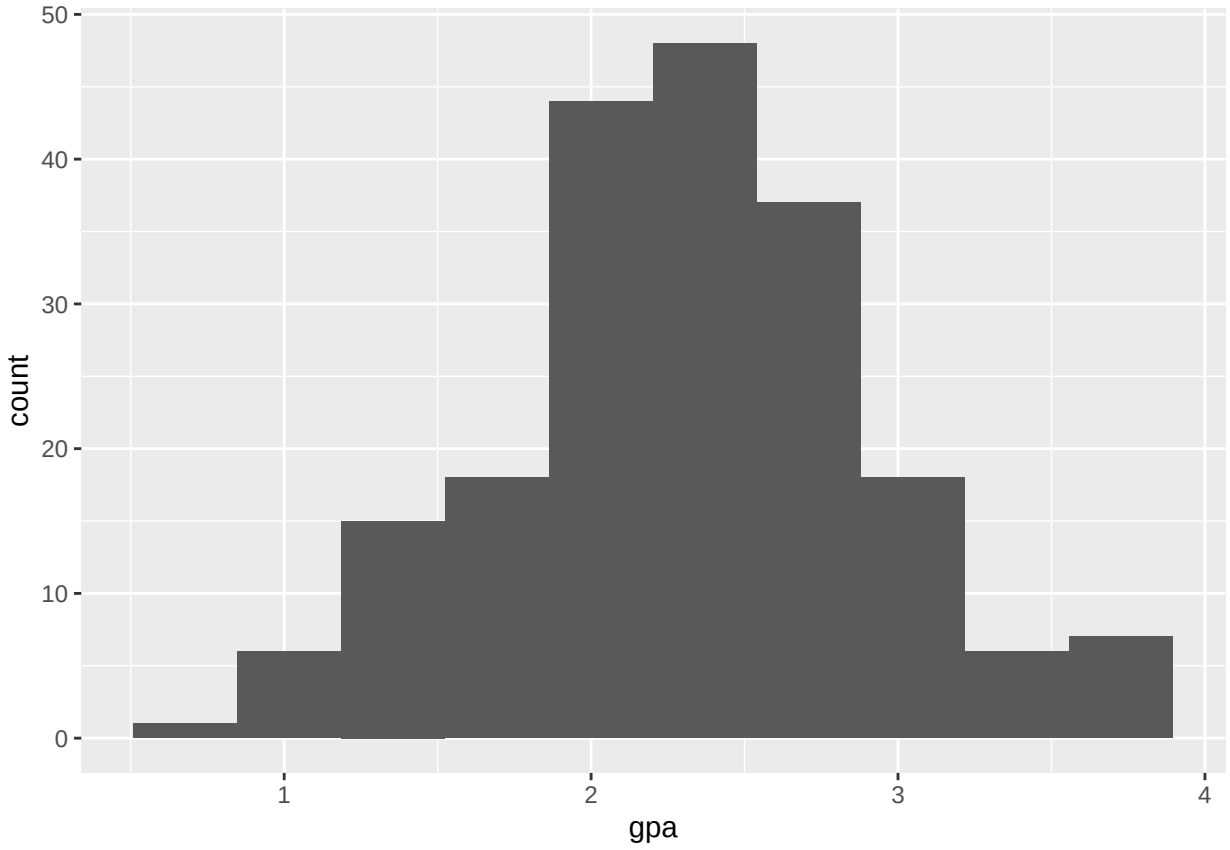
```
grades = read.table("https://www.colorado.edu/amath/sites/default/files/attached-files/grades_0.txt", s
summary(grades)
```

```
##       sat             gpa              x2                 x3
##  Min.   : 720   Min.   :0.790   Min.   :-2.86376   Min.   :1277
##  1st Qu.:1048   1st Qu.:1.965   1st Qu.:-0.71400   1st Qu.:2056
##  Median :1140   Median :2.335   Median : 0.02286   Median :2280
##  Mean   :1135   Mean   :2.319   Mean   :-0.02635   Mean   :2266
##  3rd Qu.:1240   3rd Qu.:2.683   3rd Qu.: 0.74088   3rd Qu.:2484
##  Max.   :1550   Max.   :3.840   Max.   : 2.29013   Max.   :3047
```

```
cor(grades)
```

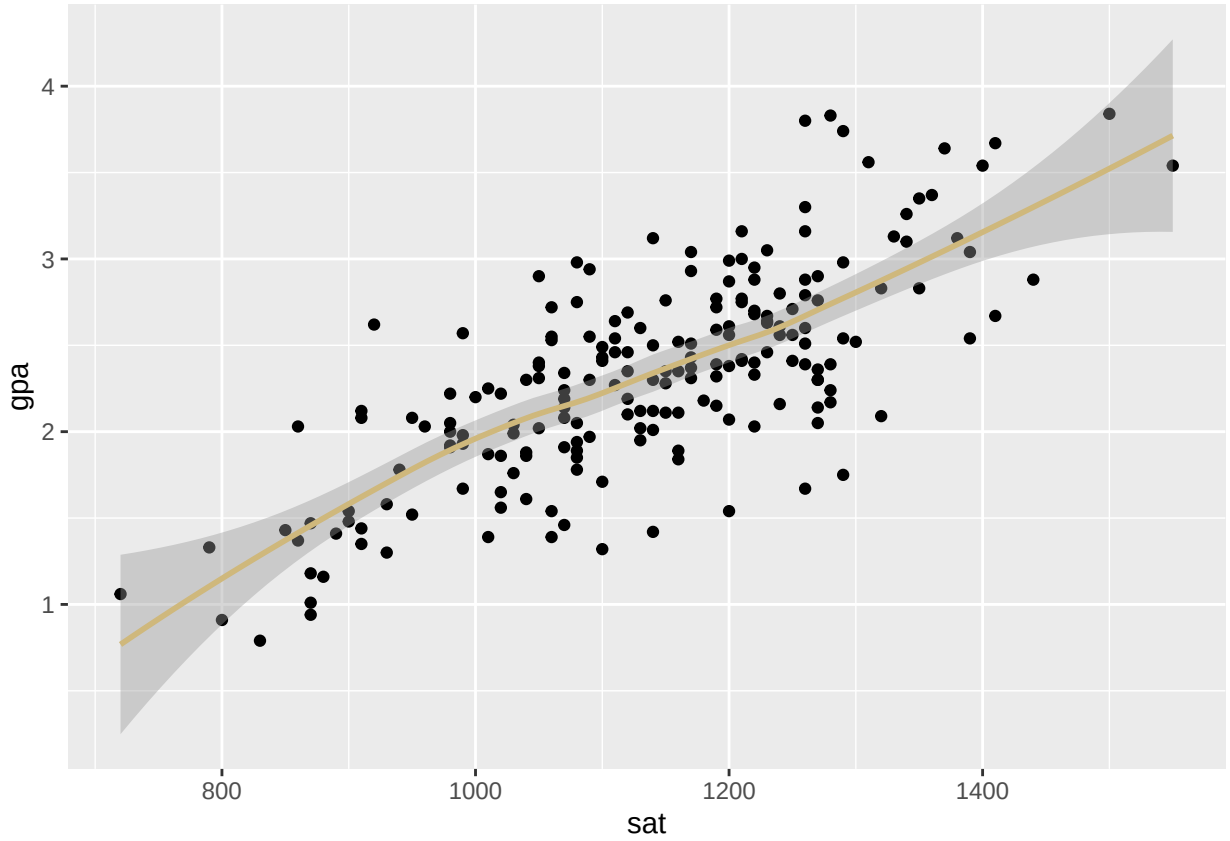
```
##            sat         gpa          x2         x3
## sat 1.00000000 0.749101520 0.023439126 0.94795613
## gpa 0.74910152 1.000000000 0.009135983 0.67025746
## x2  0.02343913 0.009135983 1.000000000 0.04361405
## x3  0.94795613 0.670257457 0.043614053 1.00000000
```

```
ggplot(grades) +
    geom_histogram(aes(x = gpa), bins = 10)
```




```
ggplot(grades) +
    geom_point(aes(x = sat, y = gpa)) +
    geom_smooth(aes(x = sat, y = gpa), col = "#CFB87C")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
lmod = lm(gpa ~ sat, data = grades)
summary(lmod)
```

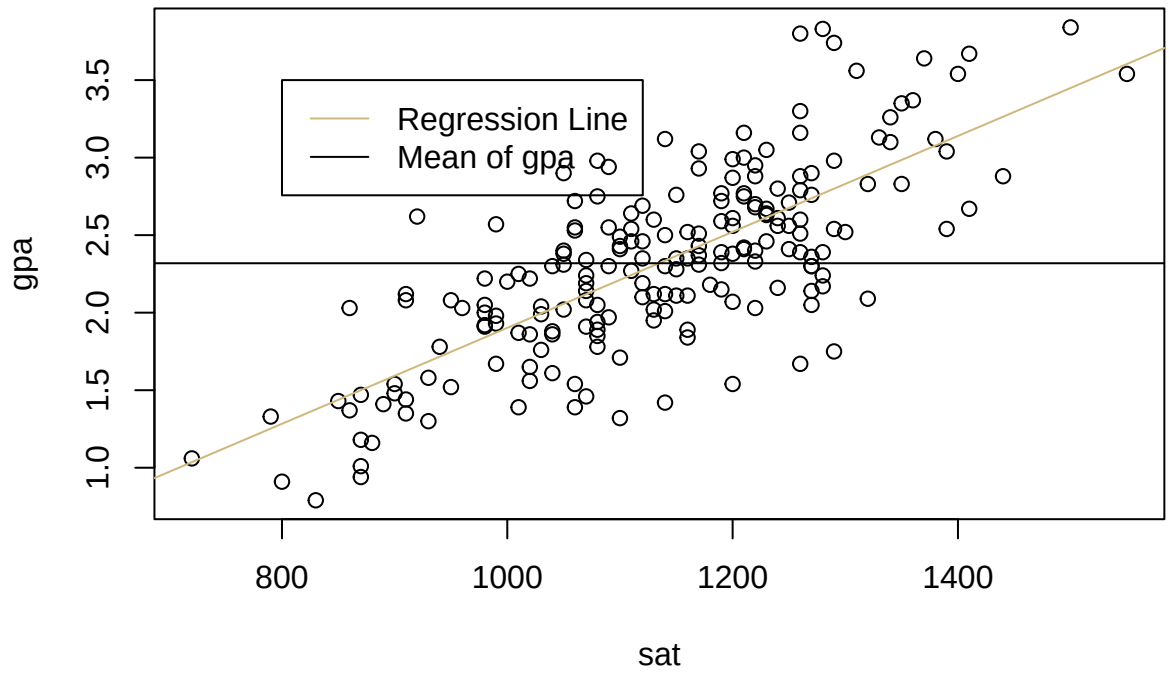
```
##
## Call:
## lm(formula = gpa ~ sat, data = grades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04954 -0.25960 -0.00655  0.26044  1.09328
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.1920638  0.2224502  -5.359 2.32e-07 ***
## sat          0.0030943  0.0001945  15.912  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3994 on 198 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5589
## F-statistic: 253.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
X = model.matrix(lmod);
(solve(t(X)%*%X))%*%t(X)%*%grades$gpa
```

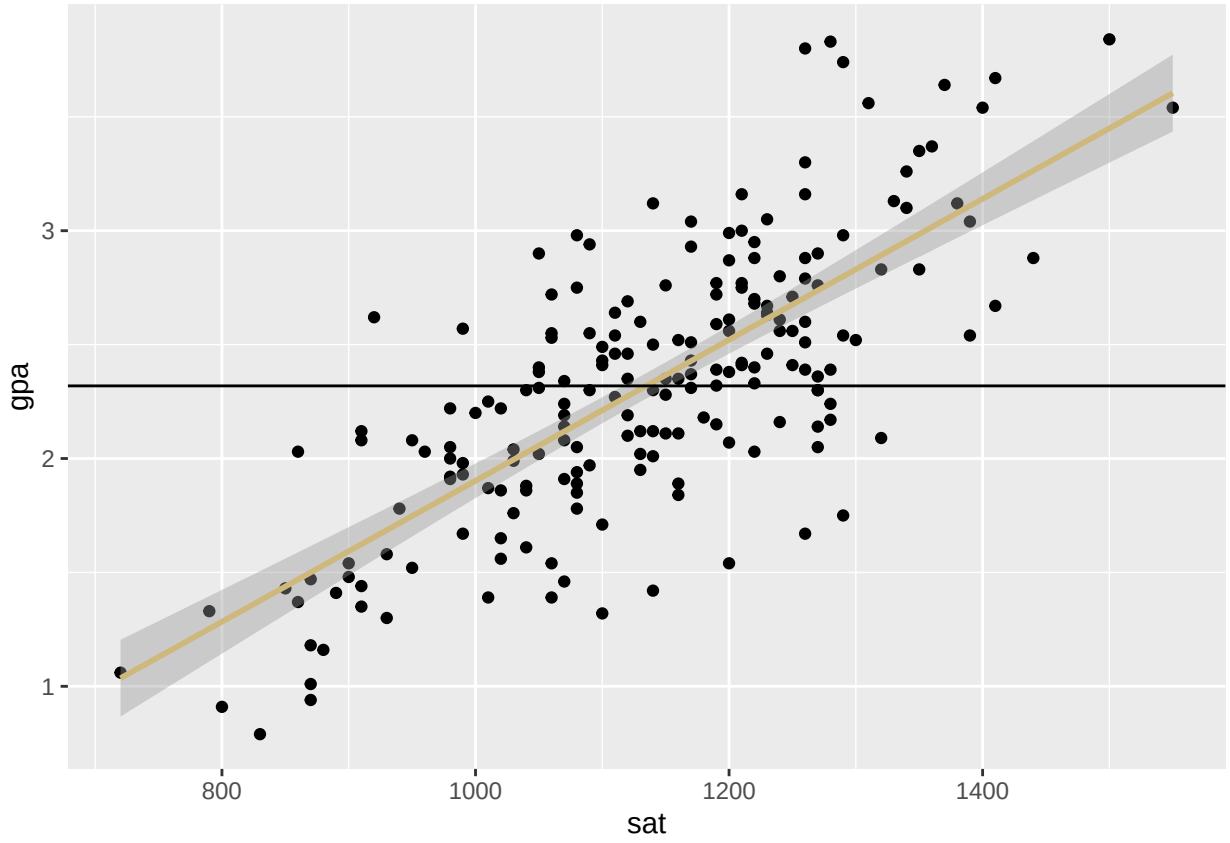
```
##                     [,1]
## (Intercept) -1.19206381
```

```
## sat          0.00309427
```

```
with(grades, plot(sat, gpa))
abline(lmod, col = "#CFB87C")
abline(h = mean(grades$gpa))
legend(800,3.5, legend = c("Regression Line", "Mean of gpa"), lty = 1, col = c("#CFB87C", "black"))
```



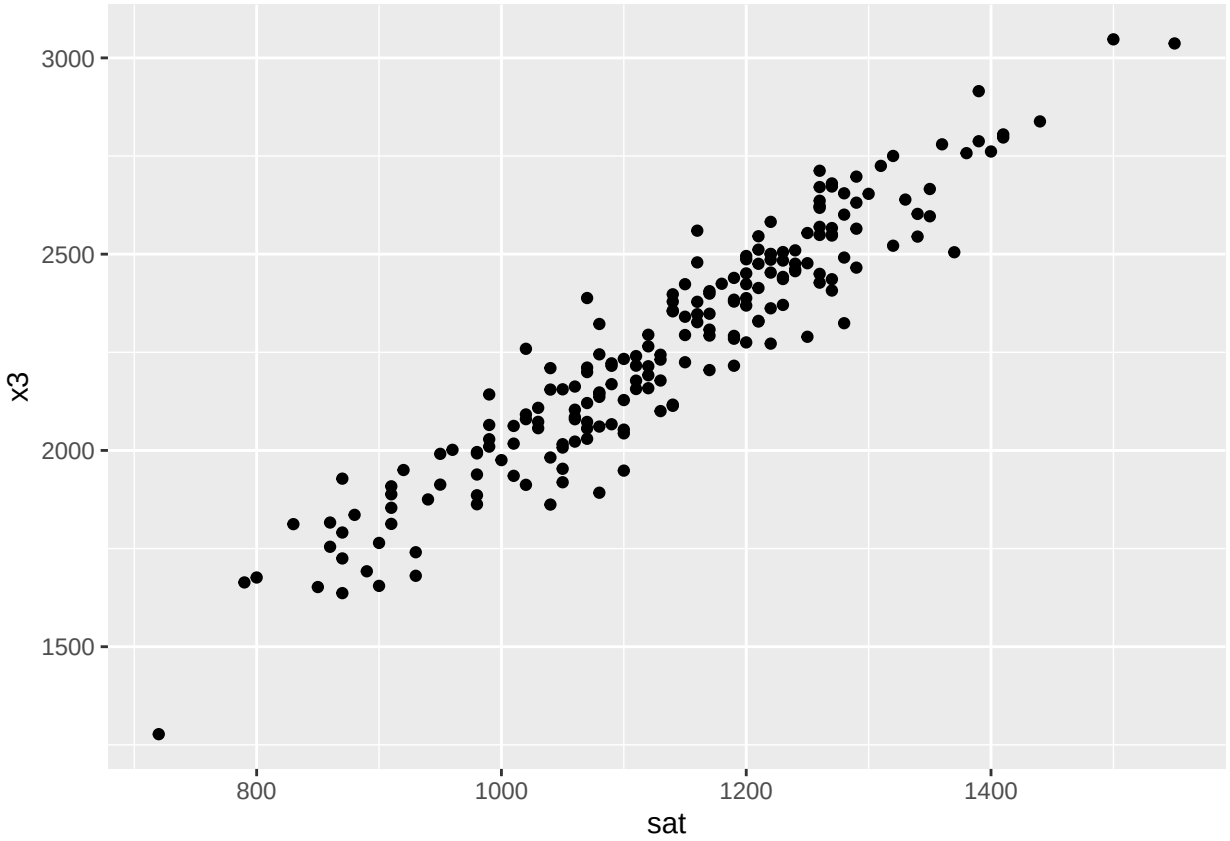

```
p = ggplot(data = grades) + geom_point(mapping = aes(x = sat, y = gpa))
p = p + geom_smooth(data=grades, formula=y~x, method=lm, color= "#CFB87C",aes(x = sat, y = gpa,group=1)
p = p + geom_hline(mapping = aes(yintercept = mean(gpa)))
p
```




```
lmod2 = lm(gpa ~ sat + x2, data = grades)
summary(lmod2)
```

```
##
## Call:
## lm(formula = gpa ~ sat + x2, data = grades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05578 -0.26424 -0.00517  0.26388  1.08778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.193114   0.223074  -5.349 2.45e-07 ***
## sat          0.003095   0.000195  15.873  < 2e-16 ***
## x2          -0.004730   0.026497  -0.179    0.859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4004 on 197 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5568
## F-statistic:   126 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
ggplot(grades) +
    geom_point(aes(x = sat, y = x3))
```




```
lmod3 = lm(gpa ~ sat + x2 + x3, data = grades)
summary(lmod3)
```

```
##
## Call:
## lm(formula = gpa ~ sat + x2 + x3, data = grades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00946 -0.24533  0.02074  0.23306  1.05049
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.178e+00  2.197e-01  -5.360 2.32e-07 ***
## sat          4.634e-03  6.038e-04   7.675 7.64e-13 ***
## x2          -6.088e-06  2.615e-02   0.000  0.99981
## x3          -7.773e-04  2.892e-04  -2.688  0.00781 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3942 on 196 degrees of freedom
## Multiple R-squared:  0.5768, Adjusted R-squared:  0.5703
## F-statistic: 89.05 on 3 and 196 DF,  p-value: < 2.2e-16
```

Let's see some ANOVA

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.1.2
```

```
data(gala, package="faraway")
gala$Species
```

```
##  [1]  58  31   3  25   2  18  24  10   8   2  97  93  58   5  40 347  51   2 104
## [20] 108  12  70 280 237 444  62 285  44  16  21
```

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz +
             Adjacent, gala)
nullmod <- lm(Species ~ 1, gala)
anova(nullmod, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ 1
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 381081
## 2     24  89231  5    291850 15.699 6.838e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmods <- lm(Species ~ Elevation + Nearest + Scruz +
              Adjacent, gala)
anova(lmods, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation + Nearest + Scruz + Adjacent
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     25 93469
## 2     24 89231  1    4237.7 1.1398 0.2963
```

```
summary(lmod)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```